

Disclaimer: This is a machine generated PDF of selected content from our databases. This functionality is provided solely for your convenience and is in no way intended to replace original scanned PDF. Neither Cengage Learning nor its licensors make any representations or warranties with respect to the machine generated PDF. The PDF is automatically generated "AS IS" and "AS AVAILABLE" and are not retained in our systems. CENGAGE LEARNING AND ITS LICENSORS SPECIFICALLY DISCLAIM ANY AND ALL EXPRESS OR IMPLIED WARRANTIES, INCLUDING WITHOUT LIMITATION, ANY WARRANTIES FOR AVAILABILITY, ACCURACY, TIMELINESS, COMPLETENESS, NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Your use of the machine generated PDF is subject to all use restrictions contained in The Cengage Learning Subscription and License Agreement and/or the Gale OneFile: Health and Medicine Terms and Conditions and by using the machine generated PDF functionality you agree to forgo any and all claims against Cengage Learning or its licensors for your use of the machine generated PDF functionality and any output derived therefrom.

Genome-Wide Functional Analysis of the Cotton Transcriptome by Creating an Integrated EST Database

Authors: Fuliang Xie, Guiling Sun, John W. Stiller and Baohong Zhang

Date: Nov. 8, 2011

From: PLoS ONE(Vol. 6, Issue 11)

Publisher: Public Library of Science

Document Type: Article

Length: 5,845 words

DOI: <http://dx.doi.org/10.1371/journal.pone.0026980>

Author(s): Fuliang Xie, Guiling Sun, John W. Stiller, Baohong Zhang *

Introduction

Cotton is among most important crops for natural textile fiber oilseed and is planted widely in 70 developed and developing countries, including the U.S., China, India, and Australia [1], [2]. Although there are more than 50 species in the genus *Gossypium*, only four of them are cultivated; these are upland cotton (*Gossypium hirsutum* L.), sea-island cotton (*Gossypium barbadense*), Asian cotton (*Gossypium arboreum*), and Arabian cotton (*Gossypium herbaceum*). Upland cotton is, by far, the most widely planted, accounting for more than 95% of the annual cotton crop worldwide.

There are approximately 45 diploid ($2n = 2x = 26$) and five tetraploid ($2n = 4x = 52$) *Gossypium* species. Upland cotton has a complex allotetraploid genome (AADD, $2n = 4x = 52$) [3], with a haploid genome size estimated to be around 2.5 Gb [4]. Decoding the cotton genome is a crucial foundation for enhancing research on fiber development, quality, yield, and other important agronomic traits. Although some progress has been made on cotton genetics and agronomic improvement, sequencing of the complete cotton genome is still ongoing, largely because of its overall genetic and structural complexity [3].

Currently, there are several types of cotton genomic resources available, including bacterial artificial chromosomes (BACs), expressed sequence tags (ESTs), linkage maps, and integrated genetic and physical maps [3]. To date, a total of 268,786 ESTs have been deposited in the public database GenBank. This large number of ESTs provides at least three obvious advantages: 1) broad EST coverage is a key landmark for future genome analysis and assembly [5]; 2) ESTs can contribute to more efficient gene discovery and identification, especially from species with unavailable genome sequences [6]; 3) ESTs provide information about gene expression, including tissue- and developmentally specific differences, as well as temporal responses to environmental changes [2]. Udall and co-workers previously assembled cotton ESTs using a total of 185,198 sequence reads from 30 cDNA libraries [7]; however, it now is necessary to re-assemble cotton ESTs because there currently are 268,786 EST reads available. Furthermore, careful investigation of the likely functions of these assembled ESTs will be more important for enhancing cotton molecular genetics, for example, identifying useful new genetic markers.

One example of such genetic markers is simple sequence repeats (SSRs), also termed microsatellites, which are tandem repeats of two-to-six base-pair nucleotide motifs. They vary in length among different genotypes and offer a rich source of

allelic polymorphisms. In contrast, SSR flanking sequences are often relatively conserved among genomes, making it possible to develop genetic markers for molecular breeding selection and genotype identification [8]-[10]. Compared with other types of molecular markers, SSRs have a number of advantages including co-dominant inheritance, high abundance, a generally random distribution across the genome, high information content, and reproducibility [9]. There are two classes of SSRs, those located in non-coding genomic regions and those found in ESTs. EST-SSRs generally are more conserved within and across related species and show higher transferability because more variable intron or intergenic sequences are absent from ESTs [11]. Additionally, it is more likely that EST-SSRs are tightly linked to specific gene functions and perhaps some even play a direct role in controlling important agronomic traits [12]. Therefore, EST-SSRs are good tools to facilitate marker-assisted selection (MAS) for breeding. To date, EST-SSRs have been used to screen cotton fiber-related loci from EST libraries generated from the cultivated diploid species *Gossypium arboreum* L. cv AKA8401 [13].

Although it is possible to find polymorphic loci using EST-SSR markers, alone they are not sufficient for uncovering the underlying genetics of highly complex traits, such as disease resistance, yield, and quality, because of their low density of coverage across the genome. Furthermore, there are limited polymorphic SSR markers available to help in discriminating between closely related species [14]. Single nucleotide polymorphisms (SNPs) are the most abundant type of DNA polymorphism in genomes. SNPs are alternative nucleotides present at a given, defined genetic location at a frequency exceeding 1% in a given population. Theoretically, each SNP can have four alleles, but bi-allelic variation has been shown to be the most frequent [15]. SNPs are considered to be the major genetic source of phenotypic variability that differentiates individuals within given species [16]. They have been applied extensively to genome-wide association studies (GWAS) of complex traits [16], fine mapping of QTLs [17], and linkage disequilibrium-based association mapping [18]. Because ESTs are rich in current public databases, it is possible for EST-derived SNPs to be a low-cost and efficient resource for investigating genome-level variability before a draft cotton genome becomes available [14], [19].

MicroRNAs (miRNAs) are short non-coding RNA molecules that regulate protein-encoding gene expression at post-transcriptional levels. The main mechanisms of miRNA action are 1) promoting degradation and 2) inhibiting translation of their target mRNAs [20]. Recently, several investigations have shown that translational inhibition is widespread in the plant kingdom [20], [21]. In plants, primary miRNAs (pri-miRNA) are transcribed by RNA polymerase II from intergenic or intron regions and then folded into pre-miRNA hairpins. DICER-LIKE 1 (DCL1) directs conversion of pri-miRNAs to pre-miRNAs, and their processing into mature miRNAs. These steps mostly are carried out in the nucleus. Mature miRNA duplexes are stabilized by the S-adenosyl methionine-dependent methyltransferase Hua Enhancer 1 (HEN1) and are exported to the cytoplasm with the assistance of the plant homolog of exportin-5, HASTY [22]. Mature miRNAs are generated by unbinding mature miRNA duplexes and then are loaded into the miRNA-induced silencing complex (miRISC). Integrated miRISC acts on a target message by perfect or near-perfect complementary base-pairing [22]. In both plants and animals, many miRNA families are highly conserved through hundreds of million of years of evolution [20]. To date, miRNAs have been identified successfully from plant EST and GSS databases based on sequence conservation and characteristic miRNA features [2], [23], [24]. EST databases also provide evidence on temporal and developmental patterns of miRNA expression. ESTs are considered to be a reliable data source for prediction of miRNAs as well their targets, especially in those species without complete genome information [2], [23], [24].

In this study, we performed global assembly of cotton ESTs available from NCBI, and functional annotation using BLASTx, BLASTn, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) resources. Using the contigs obtained, we also performed EST-based investigations of comparative transcriptome similarity between cotton and other plant species, sequence polymorphisms, expressed miRNAs and their targets, and SSR analysis. Finally, we integrated these analytical data into a comprehensive web-based database so that EST-related information can be shared and queried publically.

Results and Discussion

EST assembly

A total of 268,786 cotton ESTs were collected from NCBI; they have been obtained from different tissues, including fiber, ovule, anther, boll, callus, cotyledon, embryo, leaf, root, stem, seedling, and cultured cells (Table 1). The largest fraction of cotton ESTs is from fiber, with 114,167 sequences or 42.48% of all ESTs available. These ESTs were isolated from different treatments, including cold, cycloheximide, drought, aging, and *Fusariumoxysporum f. sp. vasinfectum* and *Xanthomonascampestris* pv. *Malvacearum* infections. After pre-processing raw sequences, a total of 235,328 clean ESTs were assembled into 28,432 unique genes (contigs) including 25,371 consensus contigs and 3,061 singletons. Contig lengths ranged from 101 to 4,080 nt (Figure 1). Consensus assemblies shared a similar sequence size distribution with singletons, except that few of the latter were found among longer length contigs. Most assembled contigs fell in the ranges from 500 nt to 900 nt (46.44%) or 900 nt to 1300 nt (26.76%) in length (Figure 1).

Figure 1. Sequence size distribution of consensus contigs and singletons in cotton. [see PDF for image]

Annotation

Because a complete cotton genome is unavailable, it is difficult to determine precise CDS and protein sequences. Gene functions were annotated in two ways: BLASTx against all plant reference proteins data and BLASTn against all plant reference nucleotide data. Most ESTs were inferred to be homologous with at least one protein-coding gene counterpart in another plant species, including *Arabidopsis*, rice, maize or grape. However, 6,441 sequences (22.64% of assembled EST contigs and singletons) by BLASTx and 7,992 contigs by BLASTn (Table 2). In total, 4,043 contigs (14.22%) could not be annotated through BLAST searches. In addition, more than 60% of ESTs shared the same or similar annotation amongst BLASTx and BLASTn search results.

The 28,432 assembled cotton contigs were further annotated by BLASTx against the GO protein database, using an E-value cutoff of $1e-20$, with 22,400 cotton ESTs finding a protein homolog (Figure 2). A total of 372 unique cellular component classes were identified for 13,657 ESTs (Figure 3A). According to annotation classification of GO database, the largest cellular component found for cotton ESTs was from cell part (6,810 contigs, 55%) and the smallest was from virion part (7 sequences, ~0%). We infer that ESTs associated with the virion part could result from contamination by virus mRNAs. A total of 13,964 ESTs were associated with 1,628 GO categories for biological processes. The majority of biological processes identified are involved in responses to stimuli (18%) and cellular process (17%) (Figure 3B). Furthermore, 15,378 ESTs were classified as involved in 1,407 molecular functions. The major molecular functions were associated with binding (57%) and catalytic (32%) activities (Figure 3C). Based on KEGG annotations from GO proteins, we made pathway enrichment analysis for cotton ESTs. This revealed 3,176 contigs to be involved in 271 different pathways (File S1).

Figure 2. Schematic pipeline for cotton EST assembly, data analysis and database development. [see PDF for image]

Figure 3. Gene Ontology (GO) analysis of 28,432 cotton annotated contigs. The three GO categories are presented: cellular component (A), biological process (B), and molecular function (C). [see PDF for image]

Using BLASTn cutoffs for E-value ([less than or equal to] $1e-30$) and sequence identity ([greater than or equal to]90%), a total of 5,461 gene clusters were identified from the entire set of 28,432 assembled cotton ESTs. The sizes of clusters varied from two to 196 members with an average size of 3.62 (Figure 4). The majority of clusters (3,358/59.8%) had 2 members.

Figure 4. Cluster size distribution of cotton contigs. [see PDF for image]

Genomic comparisons with other model plants

Based on comparisons with reference protein databases from six model species, *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Medicago truncatula*, *Oryza sativa*, *Vitis vinifera*, *Zea mays*, cotton contigs were shown to be the most similar overall to *Vitis*, followed by *Arabidopsis* (Figure 5); like cotton, both of these species are dicots. Using a BLASTx E-value cutoff of $1e-30$, 18,613 of 22,699 (82.0%) sequences from *Vitis* were found to be homologous with 19,688 of 28,432 (69.2%) cotton ESTs (Figure 5C), whereas 17,471 of 26,379 (66.2%) sequences from *Arabidopsis* were similar to 18,529 of 28,432 (65.1%) cotton contigs (Figure 5D). Amongst the six model species, *Chlamydomonas* was identified as having the least overall similarity (31.4%) to cotton. These data generally agree with current views of plant evolution; however, the highest overall similarity of cotton sequences to *Vitis* is somewhat surprising. Molecular phylogenetic analyses place the Malvaceae (cotton) and Brassicaceae (*Arabidopsis*) as sister families, with the Vitaceae (*Vitis*) a more distant outgroup [25]. The greater similarity between cotton and *Vitis* suggests that they retain somewhat more similar genome contents and sequence conservation from the common ancestor of all three taxa, than does *Arabidopsis*.

Figure 5. Homologous genomic comparison using several blast E-value cutoffs. A. Distribution of percent cotton contigs finding a hit in each genome. B. Distribution of cotton homologous proteins identified in other plant species. C. Comparison of number of homologs identified between cotton and *Vitis vinifera* with a BLASTx E-value cutoff of $1e-30$. D. The same comparison between cotton and *Arabidopsis thaliana*. [see PDF for image]

miRNAs and their targets in cotton

Because of the limited nucleotide sequence resources available, miRNA-related research in cotton has lagged far behind other plant species. Currently, only 34 cotton miRNAs have been identified and deposited into the miRBase database [26]. In this study, we used a total of 2,454 known plant miRNAs deposited in miRBase (Release 15) [26] as a reference set, and identified 87 miRNAs among cotton EST contigs and raw ESTs (Table 3). Of these, 59 were identified for the first time in cotton.

Table 3. 87 miRNAs identified in cotton ESTs. [see PDF for image]

Of the 87 miRNAs identified, 33 were from our newly assembled contigs and 54 came directly from raw EST reads (Table 3). The length of the cotton miRNAs varied from 18 to 24 nt, with average of 20.3 ± 1.4 nt (Figure 6A). The most abundant cotton miRNAs were 21 nt in length. These results are similar to miRNA lengths reported previously in plants [27]. The 87 miRNAs from cotton clustered into 57 families. The size of miRNA families in cotton varied from one to six sequence members (Table 3); 44 of 57 (77.2%) families had only one member (e.g., miR159, miR162, miR166, miR171, miR172, miR390, miR393, and miR395), whereas 13 (22.8%) had multiple members (e.g., miR156, miR164, miR394, miR398, miR399, miR414, and miR482) (Figure 6B). The largest miRNA families, including miRNA156, miRNA414, and miRNA1533, each with six members. Thirty-two of 87 miRNAs in cotton were obtained from the antisense strand of our original contig or EST, and the other 55 came from the sense strand (Table 3). miRNAs are located at either the 5' or 3' end of the hairpin arm. Our results show 50 of 87 miRNAs to be located at the 3' end and 37 at the 5' end.

Figure 6 A. Distribution of length of miRNAs in cotton. B. Size distribution of cotton miRNA families with more than one member. [see PDF for image]

Given that miRNAs target the transcripts of protein-encoding genes, a total of 18,621 ESTs, with E-values of less than $1e-25$ in BLASTx searches against the plant protein database, were selected as a subject dataset for target prediction. Based on a discrete set of criteria (see experimental procedures), 87 miRNAs identified in cotton were found to target a total of 3,260 protein-encoding genes (File S2). Our target prediction suggests that cotton miRNAs regulate the expression of many types of genes associated with diverse biological and metabolic processes, including metabolic pathways, hormone signal transduction, stress response, and fiber development. As in previous investigations, validated miRNA-target pairs also were identified in cotton, including miR156-squamosa promoter-binding protein (SBP) [28], miR164-NAC domain protein (NAC) [29], miR398-Cu/Zn superoxide dismutase [30], miR172-AP2 domain-containing transcription factor [31], and miR393-transport inhibitor response 1 [28]. In addition, because cotton is one of most important fiber crops, we also carefully examined targets associated with fiber development or fiber yield. Amongst the potential miRNA targets identified in cotton, there were at least 23 genes tightly associated with fiber development (Table 4). These targets control cellulose synthesis (miR156g and contig16368), fiber development (miR414b and contig7645), and glucose metabolism (miR529a and contig16806).

Sequence polymorphisms

We detected a total of 149,614 putative SNPs in 14,516 cotton contigs and 27,956 putative insertions/deletions (indels) in 8,674 contigs. Both SNPs and indels were detected in a total of 8,118 contigs. Our results show that SNPs occur once every 215 nt in cotton ESTs and indels occur once every 1,111 nt. The maximum frequencies of SNP and indels were 0.122 and 0.069 respectively. We generated a standard normal distribution to analyze the frequencies of SNPs/indels among contigs, and determine which contigs had a significantly high number of SNPs at $P < 0.05$ (significant) and $P < 0.01$ (highly significant). We found 1,933 contigs to contain significant SNP frequencies, with 802 of these contigs at high significance. A significant frequency of indels was found for 1,089 contigs, 735 of which were highly significant. Currently, the genome of cotton is incompletely sequenced; in its absence, however, the large resource of ESTs available allow for identification of large numbers of SNPs [14]. The apparently high frequency of SNPs and indels we observed in cotton ESTs could be due in part to sequencing errors. To address this issue, we followed the criteria of Wang and co-workers [14] to remove pseudo-SNPs and pseudo-indels as much as possible. Without experimental validation, however, it is difficult to determine whether a given SNP or an indel in cotton represents a real polymorphism. Nevertheless, we suggest that the high average frequency of SNPs we observed could, indeed, reflect real genetic variation resulting from the complicated genetic background present in large cotton EST libraries. However, because of the nature of cotton EST data in the NCBI database, it is not 100% sure that these SNPs are really SNPs or caused by sequencing errors. As deep sequencing technology become available, more study may be performed to investigate this issue.

Aside from those that could not be assigned a presumed function, many cotton EST contigs with significant rate of SNPs and indels are associated with transcription factors, energy metabolism, stress response, signal transduction, and protein kinases (File S3). A previous investigation showed that high SNP frequency (0.013) occurred in R2R3-MYB transcription factors from cotton [32]. In this study, we also detected two contigs (contig2733 and contig15263) annotated to encode MYB transcription factors that have significantly high SNP frequencies. Therefore, it is possible that the high diversity of SNPs and indels in the cotton transcriptome could be related to functional adaptations to environmental stress.

Simple sequence repeats

Because of their relative abundance and ease of generation, SSRs are among the most powerful of molecular markers, and have been applied widely in molecular-assisted selection (MAS) for plant breeding programs [33]. SSR markers derived from expressed sequence tags (EST-SSRs) originate from transcribed regions of the genome and are likely to be even more transferable across lines, populations and species than random genomic SSRs [13]. In this study, we analyzed SSRs in both cotton contigs and raw ESTs. We identified a total of 151 SSRs from cotton contigs and 4,214 from raw ESTs (File S4). Among SSRs from contigs, the most abundant repeat types were trinucleotides (130, 86.09%) followed by dinucleotides (21, 13.91%). The dominant sequence repeat in contigs was AAG/CTT (10, 6.62%) followed by TGA/TCA (9, 5.96%). Trinucleotide repeats also were the most common among SSRs from raw ESTs (2,961, 70.27%) again followed by dinucleotides (829, 19.67%)

along with a sizeable fraction of tetranucleotides (424, 10.06%). Dominant repeat types in raw ESTs were GAA/TTC (159, 3.77%) and GAT/ATC (159, 3.77%). Amongst the 151 SSR markers found, only 43 come from the contigs annotated with known functions. Potentially, these markers could be exploited for use in marker-assist breeding selection. Of these SSRs, 51 from contigs and 1,663 from raw ESTs have not been reported previously in cotton.

In further investigate the potential of these SSR repeats as genetic markers, we employed eprimer3 (primer 3) to design primer pairs for each SSR under a series of primer-designing parameters (see Experimental procedures). We were able to find viable primer pairs for 121 of 151 contig SSRs and 3,092 of 4,214 raw EST SSRs (all these primers can be downloaded from the cotton EST website www.leonxie.com).

Web-based database for cotton ESTs

To facilitate further investigation and application of cotton genome-related research, we constructed a web-based, searchable and downloadable database for managing cotton ESTs data, along with related deep sequence analyses including assembly, annotation, miRNAs, SNP and indels, and SSRs (Figure 2). This database can be accessed freely through a web interface (www.leonxie.com). Raw ESTs, as well as annotation and assembly data can be queried using different strategies, such as gene accession, gene ID, and function (Figure 7). We also incorporated the Cotton Marker Database (CMD) into our web-server and built connections with raw EST, assembled contigs, and SSR databases. In this way, users can quickly access marker information from cotton ESTs or access marker-related ESTs through CMD markers. We have attempted to develop a seamless connection among all of these cotton EST datasets and resources. For instance, when investigating a contig, users can visit its related information, including functional annotation, miRNA, SSR, SNP, GO, and KEGG; alternatively that contig can be accessed from any one of the related resources as a starting point. To improve the efficiency of BLAST analyses of cotton ESTs, we also built a local WWW-BLAST server permitting directed and advanced BLAST options. Raw cotton ESTs, assembled contigs, consensus assemblies, singletons, all reference protein databases from plants, and all reference plant nucleotide databases are incorporated within our local WWW-BLAST server as potential query targets. Furthermore, EST data and related analytical tools and results, all can be freely accessed and downloaded.

Figure 7. Interface of cotton EST database for querying raw ESTs (A), and assembled contigs (B). [see PDF for image]

Conclusions

We have developed a specific and dedicated workbench for assembling cotton ESTs and for performing genome-wide analyses of the cotton transcriptome. In addition to raw ESTs and assembled contigs, additional EST-related information, including miRNAs, SNPs, and SSRs has been integrated into this database. A friendly web-interface allows users to access and download these data as batch files or via directed searches based on specific interests and needs. Moreover, now that this platform for cotton EST data has been established, it will be very convenient to add new cotton ESTs and annotated resources to our database in future. Therefore, this cotton EST database can contribute significantly to advancing research on cotton ESTs and global genome-wide analyses.

Methods

Dataset

A total of 268,786 cotton ESTs (*Gossypium hirsutum* L.) were downloaded from NCBI (<http://www.ncbi.nlm.nih.gov/>). These ESTs were obtained from at least 90 EST libraries and samples treated under at least eight different abiotic and biotic conditions.

Data pre-processing

A majority of raw EST sequences potentially contain various contaminating elements, such as sequencing primers, vector sequence, sequences from other species, and sequencing errors. In addition, poly A/T tail and low complexity sequences are inevitably present in some raw ESTs. Thus, a critical first step is to remove these contaminated sequences before performing more deep analysis. In this study, we first cleaned original cotton ESTs by Seqclean [34] (<ftp://ftp.tigr.org/pub/software/tgi/seqclean/>) from TIGR under default parameters. Seqclean is a versatile tool for removing sequences from vectors, mitochondria, ribosomal RNAs, sequencing primers, polyA/T tails, low complexity sequences, and sequences with lengths under 100 nt [34]. After processing with SeqClean (Figure 2), we employed RepeatMasker (version 3.2.9, <http://www.repeatmasker.org/>) to mask repeated elements based on Repbase (Repbase 15.04, <http://www.girinst.org/>) [35]. Finally, a total of 235,328 cleaned ESTs were kept for further assembly.

EST clustering and assembling

The cleaned EST sequences were clustered and assembled into contigs (consensus and singletons) by TGICL (<http://ftp.tigr.org/pub/software/tgi/tgicl/>) [36], which could partition the input dataset into small groups of sequences (clusters) using Megablast and assemble each cluster by using the cap3 program [37] into contigs. The resulted data was further performed an ortholog search against the published assembled data of *Gossypium*'s ESTs (<http://www.agcol.arizona.edu/cgi-bin/pave/Cotton/index.cgi>) [7] using Orthomcl (Version 2.0, <http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=orthomcl#Software>) under the cutoff of E-value of 1e-25 and identify of 95%.

Functional annotation

In order to investigate putative functions of cotton ESTs, we performed BLASTx [38] against reference protein databases from all plants using an E-value cutoff of 1e-20, and BLASTn against reference nucleotide acid databases from all plants at an E-value cutoff of 1e-25. Only the best high-scoring segment pair (HSP) was kept for annotation. We also tried to annotate possible open reading frames (ORFs) of contigs and further infer their protein sequences by GETORF from Emboss tools package (<http://emboss.sourceforge.net/>). The longest ORF was considered to be the candidate CDS sequence, and its translation the presumed protein sequence as well.

To better understand the functional classification of ESTs, contigs were used as queries in BLASTx using Gene Ontology (GO) analysis [39]. Cellular component, biological process, and molecular function were classified for these contigs. We performed further pathway enrichment according to GO annotations for Kyoto Encyclopedia of Genes and Genomes (KEGG) [40].

Cluster analysis

Each individual contig was queried against the complete assembled EST data set using BLASTn. All contigs hit by the query with an E-value of less than 1e-30 and an identity of more than 90% were defined as a cluster.

Overall genomic sequence similarity

Using different BLASTx E-value cutoffs (E[less than or equal to]1e-10, E[less than or equal to]1e-30, E[less than or equal to]1e-50, and E[less than or equal to]1e-100), we investigated sequence similarity between the cotton contigs we obtained and reference cDNA databases from several model species; these included *Arabidopsis thaliana* (TAIR9, http://ftp.arabidopsis.org/Sequences/blast_datasets/TAIR9_blastsets/), *Chlamydomonas reinhardtii* (Chlre4, <http://genome.jgi-psf.org/chlamy/chlamy.download.ftp.html>), *Medicago truncatula* (Mt3.0 release, <http://www.medicago.org/genome/downloads.php>), *Vitis vinifera* (http://ftp.ncbi.nih.gov/genomes/Vitis_vinifera/Assembled_chromosomes/), *Zea mays* (<http://www.plantgdb.org/ZmGDB/cgi-bin/downloadGDB.pl>), and *Oryza Sativa* (version 6.1, http://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_6.1/all.dir/).

Sequence polymorphism analysis

Based on assembly results of consensus contigs, SNP and indel polymorphisms were analyzed. A perl script was developed to detect SNPs and indels under several criteria as described by Wang and co-workers [14]. Briefly, 1) a mismatch identified within contigs containing more than four individual EST reads was definable as a SNP or an indel; 2) variation among sequences was considered to be a bona fide SNP or indel polymorphism when it was found at least twice within contigs assembled by 5-6 ESTs; 3) at least three times within contigs assembled by 7-8 ESTs; 4) at least four times within contigs assembled by 9-12 ESTs; 5) and at least five times within contigs assembled by 13 or more ESTs.

Identification of miRNAs and their targets

MicroRNAs (miRNAs) are known as a class of non-coding endogenous small RNA molecules with lengths of ~21 nt. Investigations increasingly show that miRNAs regulate target mRNAs either by inducing their degradation or by inhibiting translation [20]. To date, miRNAs have been predicted successfully from various EST [41] and GSS databases [23]. Especially for those species without complete genome information, an EST database is considered to be an ideal data source for predicting miRNAs their targets as well [24], [42]. In our analysis, low complexity sequences, sequences with lengths of less than 100 nt, and sequences with repeated elements were removed in data pre-processing; EST contigs generated and raw ESTs then were combined as the subject dataset. We employed all known plant miRNAs from miRBase (Release 15: April 2010, <http://www.mirbase.org/>) [26] as a reference set and performed homology searches against the subject dataset using methods reported previously [43], [44]. Cotton miRNA targets also were predicted according to method in previous reports [43].

SSR detection and primer design

In order to locate simple sequence repeats (SSRs) in cotton ESTs, we performed SSR analyses on cotton contigs and raw ESTs using a software SSR Finder from GRAMENE (<ftp://ftp.gramene.org/pub/gramene/software/scripts/ssr.pl>). The parameters were designed for identifying perfect di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of 6, 5, 4, 4, and 4 repeats respectively [9]. Eprimer3 from EMBOSS bioinformatics software packages (<http://emboss.sourceforge.net/>) [45] was used to design flanking primers for detected microsatellites. The major parameters for primer design were set as following: PCR products ranging from 100 to 300 nt; primer lengths ranging from 18 to 24 nt with an optimum of 20 nt, 60°C optimal annealing temperature, and GC content from 40%~65% with an optimum of 50% [9].

Construction of a web-based cotton EST database

In order to share our integrated data and analytical results on cotton ESTs, including raw ESTs, assembled EST contigs, predicted miRNAs, sequence polymorphisms, and SSRs and primers, we integrated the information from each step of our investigation into a web-based cotton EST database, using open-source software (Apache, PHP, and MySQL), and constructed interfaces among the data types (Figure 2). Furthermore, to facilitate access to potentially useful markers from cotton raw ESTs and assembled contigs, we incorporated current data (SSR and QTL) from the Cotton Marker Database (CMD) (<http://www.cottonmarker.org/>) into our EST database. Our new web-based cotton EST database provides users with a friendly interface to query or download data. It is freely available at the website www.leonxie.com.

Supporting Information

File S1 Pathway analysis by KEGG.

(XLS)

File S2 Predicted miRNA targets.

(XLS)

File S3 Cotton EST contigs with significant SNPs and indels.

(XLS)

File S4 Identified SSR markers with designed primers.

(XLS)

Author Contributions

Conceived and designed the experiments: FX GS JS BZ. Performed the experiments: FX GS BZ. Analyzed the data: FX GS JS BZ. Contributed reagents/materials/analysis tools: FX JS BZ. Wrote the paper: FX GS JS BZ.

References

1. IAC. 1996 Cotton: Review of World Situation, Monogram by International Advisory Committee. Washington, D.C
2. Zhang BHWang QLWang KBPan XPLiu F 2007 Identification of cotton microRNAs and their targets. Gene 397 2637
3. Chen ZJScheffler BEDennis ETriplett BAZhang T 2007 Toward sequencing cotton (*Gossypium*) genomes. Plant Physiol 145 13031310

4. Hendrix B Stewart JM 2005 Estimation of the nuclear DNA content of gossypium species. *Ann Bot* 95 789797
5. Seki M Hayashida N Kato N Yohda M Shinozaki K 1997 Rapid construction of a transcription map for a cosmid contig of *Arabidopsis thaliana* genome using a novel cDNA selection method. *Plant J* 12 481487
6. Hattori J Ouellet T Tinker NA 2005 Wheat EST sequence assembly facilitates comparison of gene contents among plant species and discovery of novel genes. *Genome* 48 197206
7. Udall J A Swanson J M Haller K Rapp R A Sparks ME 2006 A global assembly of cotton ESTs. *Genome Res* 16 441450
8. Pearson C E Sinden RR 1998 Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Curr Opin Struct Biol* 8 321330
9. Zeng S Xiao G Guo J Fei Z Xu Y 2010 Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* 11 94
10. Sanchez de la Hoz M P Davila J A Loarce Y Ferrer E 1996 Simple sequence repeat primers used in polymerase chain reaction amplifications to study genetic diversity in barley. *Genome* 39 112117
11. Varshney R K Graner A Sorrells ME 2005 Genic microsatellite markers in plants: features and applications. *Trends Biotechnol* 23 4855
12. Bozhko M Riegel R Schubert R Muller-Starck G 2003 A cyclophilin gene marker confirming geographical differentiation of Norway spruce populations and indicating viability response on excess soil-born salinity. *Mol Ecol* 12 31473155
13. Park Y H Alabady M S Ulloa M Sickler B Wilkins TA 2005 Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. *Mol Genet Genomics* 274 428441
14. Wang S Sha Z Sonstegard T Liu H Xu P 2008 Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics* 9 450
15. Krawczak M 1999 Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis* 20 16761681
16. Nicolae D L Gamazon E Zhang W Duan S Dolan ME 2010 Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 6 e1000888
17. Zhang J Lu Y Yuan Y Zhang X Geng J 2009 Map-based cloning and characterization of a gene controlling hairiness and seed coat color traits in *Brassica rapa*. *Plant Mol Biol* 69 553563
18. Schneider K Kulosa D Soerensen T R Mohring S Heine M 2007 Analysis of DNA polymorphisms in sugar beet (*Beta vulgaris* L.) and development of an SNP-based map of expressed genes. *Theor Appl Genet* 115 601615
19. Li F Kitashiba H Inaba K Nishio T 2009 A *Brassica rapa* linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits. *DNA Res* 16 311323
20. Bartel DP 2004 MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116 281297
21. Brodersen P Sakvarelidze-Achard L Bruun-Rasmussen M Dunoyer P Yamamoto Y Y 2008 Widespread translational inhibition by plant miRNAs and siRNAs. *Science* 320 11851190
22. Voinnet O 2009 Origin, biogenesis, and activity of plant microRNAs. *Cell* 136 669687
23. Zhang B Pan X Anderson TA 2006 Identification of 188 conserved maize microRNAs and their targets. *FEBS Lett* 580

24. Xie FLHuang SQGuo KXiang ALZhu YY 2007 Computational identification of novel microRNAs and targets in *Brassica napus*. *FEBS Lett* 581 14641474
25. Wang HCMoore MJSoltis PSBell CDBrockington SF 2009 Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences of the United States of America* 106 38533858
26. Griffiths-Jones SSaini HKvan Dongen SENright AJ 2008 miRBase: tools for microRNA genomics. *Nucleic Acids Research* 36 D154D158
27. Zhang BHPan XPCox SBCobb GPAnderson TA 2006 Evidence that miRNAs are different from other RNAs. *Cellular and Molecular Life Sciences* 63 246254
28. Schwab RPalatnik JFRiester MSchommer CSchmid M 2005 Specific effects of microRNAs on the plant transcriptome. *Dev Cell* 8 517527
29. Guo HSXie QFei JFChua NH 2005 MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for arabidopsis lateral root development. *Plant Cell* 17 13761386
30. Sunkar RKapoor AZhu JK 2006 Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in *Arabidopsis* is mediated by downregulation of miR398 and important for oxidative stress tolerance. *Plant Cell* 18 20512065
31. Aukerman MJSakai H 2003 Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell* 15 27302741
32. An CSaha SJenkins JNMa DPScheffler BE 2008 Cotton (*Gossypium* spp.) R2R3-MYB transcription factors SNP identification, phylogenomic characterization, chromosome localization, and linkage mapping. *Theor Appl Genet* 116 10151026
33. Kantartzi SKUlloa MSacks EStewart JM 2009 Assessing genetic diversity in *Gossypium arboreum* L. cultivars using genomic and EST-derived microsatellites. *Genetica* 136 141147
34. Chen YALin CCWang CDWu HBHwang PI 2007 An optimized procedure greatly improves EST vector contamination removal. *Bmc Genomics* 8 11
35. Jurka J 1998 Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol* 8 333337
36. Pertea GHuang XLiang FAntonescu VSultana R 2003 TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19 651652
37. Huang XMadan A 1999 CAP3: A DNA sequence assembly program. *Genome Res* 9 868877
38. Altschul SFMadden TLSchaffer AAZhang JHZhang Z 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25 33893402
39. Ashburner MBall CABlake JABotstein DButler H 2000 Gene Ontology: tool for the unification of biology. *Nature Genetics* 25 2529
40. Kanehisa MGoto S 2000 KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28 2730
41. Zhang BPan XCannon CHCobb GPAnderson TA 2006 Conservation and divergence of plant microRNA genes. *Plant J* 46 243259

42. Venne LSAnderson TAZhang BSmith LMMcMurry ST 2008 Organochlorine pesticide concentrations in sediment and amphibian tissue in playa wetlands in the Southern High Plains, USA. *Bulletin of Environmental Contamination and Toxicology* 80 497501

43. Xie FFrazier TPZhang B 2010 Identification and characterization of microRNAs and their targets in the bioenergy plant switchgrass (*Panicum virgatum*). *Planta* 232 417434

44. Zhang BHPan XPWang QLCobb GPAnderson TA 2005 Identification and characterization of new plant microRNAs using EST analysis. *Cell Research* 15 336360

45. Rychlik W 1995 Selection of primers for polymerase chain reaction. *Mol Biotechnol* 3 129134

Author Affiliation:

Department of Biology, East Carolina University, Greenville, North Carolina, United States of America

Corresponding Author: * E-mail: zhangb@ecu.edu

Editor: Christos A. Ouzounis,

Article History:

Received Date: 7/6/2011

Accepted Date: 10/7/2011

Published Date: 11/8/2011

Copyright: © 2011 Xie et al

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported partially by the USDA and the Cotton Incorporated. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

DOI: 10.1371/journal.pone.0026980

Fuliang Xie; Guiling Sun; John W. Stiller; Baohong Zhang

Copyright: COPYRIGHT 2011 Public Library of Science
<http://www.plosone.org/static/information>

Source Citation

Xie, Fuliang, et al. "Genome-Wide Functional Analysis of the Cotton Transcriptome by Creating an Integrated EST Database." *PLoS ONE*, vol. 6, no. 11, 2011, p. e26980. *Gale OneFile: Health and Medicine*, <https://link.gale.com/apps/doc/A476863338/HRCA?u=ncliveecu&sid=HRCA&xid=55fea51a>. Accessed 2 Apr. 2020.

Gale Document Number: GALE|A476863338
